DOCUMENT RESUME

ED 462 262 SE 065 269

AUTHOR Epstein, Jonathan A.; Kans, Jonathan A.; Schuler, Gregory D.
TITLE WWW Entrez: A Hypertext Retrieval Tool for Molecular Biology.
PUB DATE 1994-10-00
NOTE 10p.; Contains small print. Paper part of the World Wide Web Conference: Mosaic and the Web (2nd, 1994).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Biological Sciences; *Database Management Systems; Databases; Higher Education; *Information Management; Internet; Research Tools; World Wide Web

ABSTRACT

This article describes the World Wide Web (WWW) Entrez server which is based upon the National Center for Biotechnology Information's (NCBI) Entrez retrieval database and software. Entrez is a molecular sequence retrieval system that contains an integrated view of portions of Medline and all publicly available nucleotide and protein databases, including GenBank. A CD-ROM is available but the WWW server enables those who access the databases to have enhanced capabilities. Sections relating to the history of Entrez and WWW Entrez, a comparison of three types of Entrez, and a discussion of conclusions and future directions are included. (DDR)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 462 262

WWW Entrez: A Hypertext Retrieval Tool for Molecular Biology.

Jonathan A. Epstein
Jonathan A. Kans
Gregory D. Schuler

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

- ☐ This document has been reproduced as received from the person or organization originating it.
- ☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. Epstein

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

SE 065 269

BEST COPY AVAILABLE

WWW Entrez: A Hypertext Retrieval Tool for Molecular Biology

Jonathan A. Epstein, Management Systems Designers, Inc.

Jonathan A. Kans, National Center for Biotechology Information

Gregory D. Schuler, National Center for Biotechnology Information

Abstract

The WWW Entrez server is a WWW interface based upon the National Center for Biotechnology Information's Entrez retrieval database and software. Entrez is a molecular sequence retrieval system, which contains an integrated view of portions of MEDLINE, and all publically available nucleotide and protein databases, including GenBank.

While Entrez was already available as both a CD-ROM and a true standalone client/server application ("Network Entrez"), a WWW server was developed to obtain the advantages of the World Wide Web. These advantages include the wide availability of WWW clients, support for "dumb" terminals which Entrez does not support, and the benefits of client software which is supported by others (e.g., NCSA Mosaic).

The availability of alternate implementations of Entrez provides us with a rare opportunity to investigate the strengths and weaknesses of the WWW. In particular, we can compare the ergonomics, performance and bandwidth requirements of WWW Entrez to those of the true client/server implementation. In addition, the relative ease of integrating the two network services into other molecular biology software is discussed. Finally, we present usage statistics for WWW Entrez, the Entrez CD-ROM, and Network Entrez; the degree to which these services are actually used may provide the ultimate measure of effectiveness.

Contents

- Abstract
- History
  - Entrez CD-ROM
  - Network Entrez
  - WWW Entrez
- Ergonomics
  - Entrez
  - WWW Entrez
- WWW Entrez discussion
- Comparison of three types of Entrez
- Usage
- Conclusions and future directions
- Bibliography

History

Like many WWW servers, the WWW Entrez server had an earlier incarnation as a retrieval system. Unlike most such systems, however, Entrez also existed as a true client/server Internet application.

Entrez CD-ROM

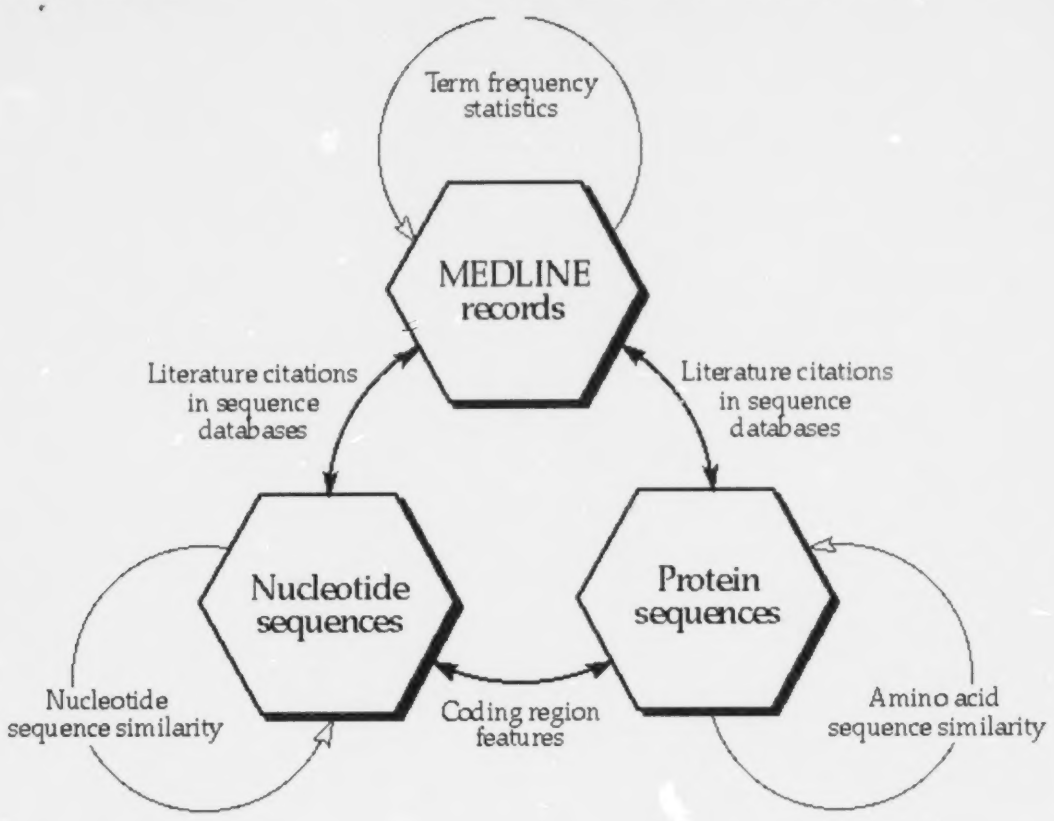
In 1992, the National Center for Biotechnology Information (NCBI) began distributing the *Entrez* CD-ROM subscription, issued six times annually.

The CD-ROM subscription, containing both data and software, provides an integrated view of the public DNA and protein databases, as well as a related subset of the MEDLINE medical literature database.

The growth of information which molecular biologists need to access to support their experimental work is growing at an explosive rate. The scientific literature adds some 6,000 peer-reviewed articles per month. Databases which contain the codes for gene sequences double in size every 20 months. Biologists need a retrieval tool which offers easy, accurate and complete access to this type of information and this was the motivation for developing Entrez. Entrez is a search tool for integrated access to the biological literature and sequence data [3].

In addition to conventional indexed lookup techniques and the ability to link among the three databases, Entrez also provides the ability to link **within** a database by finding "related" sequences or literature abstracts. These **neighboring** relationships within the nucleotide database are pre-computed by comparing the entire nucleotide database against itself using the BLAST[1] sequence similarity algorithm. Proteins are also neighbored using BLAST, and the literature abstract database is neighbored against itself using a statistical text retrieval algorithm [6].

SEQ5349



Network Entrez

A network-based retrieval mechanism for the Entrez CD-ROM data was needed since CD-ROMs faced the following limitations:

- Slow access speed
- data currency
- capacity is limited to 660 Mbytes per CD-ROM.

Because of these advantages, *Network Entrez* was written, and released in mid-1993. Network Entrez is a true client/server software package where the client and server transmit and receive data according to a well-defined Abstract Syntax Notation 1 (ASN.1[5]) protocol, and only small amounts of data are sent across the network to facilitate smooth software operation. The Network Entrez client obtains a connection with a suitable Network Entrez server by first connecting to a central *Dispatcher* at a well-known address. The use of the client< - > Dispatcher protocol, also an ASN.1 protocol, results in the Dispatcher instructing a suitable server to connect to the client. The desired client< - > server connection is thereby established.

The communications protocol provides a desirable layer of abstraction between the client and server. In practice, there are two different Network Entrez servers which use different types of underlying databases, but understand the same ASN.1 protocol.

To reduce NCBI's software support requirements and to provide local support to users, a Network Entrez administrator at each site takes responsibility for local users. As of this writing (September 1994), there are over 800 registered Network Entrez sites in 32 countries.

WWW Entrez

While Network Entrez opened access to Entrez for a large number of Internet users, there were several motivations for building a WWW version of Entrez:

- Serve vt-100 class users who previously did not have access to Entrez, since Entrez and Network-Entrez are window-based applications. Note that lynx, for example, is a WWW browser which does not require a windowed environment. [Note: a vt-100 Entrez navigator, CLEVER, was concurrently developed independently of WWW Entrez]
- The linking and neighboring information available in Entrez can be naturally expressed as hypertext
- The universality of the World Wide Web
- WWW browsers are supported by third-part developers
- The ability to link to external data sources on the Web.

The WWW Entrez server was constructed using a combination of *Bourne* shell scripts and a single C program. This C program, `entrcmd`, is a search engine which can perform powerful retrievals using a simple UNIX-command-line query language. For example, `entrcmd` can look up the entries associated with a Boolean expression, and then perform an arbitrary number of rounds of inter-database linking and intra-database neighboring using the MEDLINE, protein, and nucleotide databases. Also, given a starting term (e.g., "Jones JD") it can fetch a number of consecutive terms from a given field (e.g., 50 author names). In its role as the WWW server search engine, `entrcmd` only runs on a UNIX host; however, it is layered on top of the NCBI toolbox, which is portable across a wide variety of platforms. The source code for `entrcmd` appears within the NCBI toolbox, and can

run on all the NCBI-supported platforms.

Ergonomics

Ergonomics of Entrez

The ergonomics of the WWW Entrez server were intended to be similar to the existing Entrez programs, wherever possible. In Entrez (both CD-ROM and Network versions), a user can perform an initial Boolean query by selecting:

- a database (MEDLINE, protein, or nucleotide),
- a field (e.g., "Author Name" or "Journal Title"), **and**
- a mode which allows query terms to be viewed in different ways (e.g., "selection mode", which allows the user to scroll through all the available terms, alphabetically).

Having selected one or more terms in this manner, the user may perform simple Boolean operations by clicking and using drag-and-drop. More complex Boolean queries can be performed using a combination of point-and-click and typing.

In Entrez, each toggle of a popup (e.g., database or field) and each entry of a datum results in an action by the Entrez application program. Furthermore, continuous scrolling of huge alphabetical lists is possible, since moderate scrolling can result in the fetching of only a small number of terms. Experience shows that this scrolling ("selection mode") is a valuable feature of Entrez, since users don't always know the exact query term or author name.

Once a Boolean query has been composed by the user, the user can fetch and view the resulting document summaries. Again, these document summaries appear as a scrolled list, and only a few entries need to be retrieved at a time. The user may then view an entire entry by double-clicking on a document summary of interest.

Links to other databases or neighboring within the current database may be performed by selecting the target database and marking one or more articles to be used as the basis for linking/neighboring. For example, by marking several MEDLINE articles a user can retrieve the related sequence data from the DNA database.

Ergonomics of WWW Entrez

In WWW Entrez, an attempt has been made to preserve as much of the familiar look of Entrez as possible, without wasting unreasonable amounts of network bandwidth. WWW Entrez is inherently slower than Network Entrez. This is because:

- The portion of the WWW Entrez server which uses the `entrcmd` engine is written using Bourne shell scripts. If it were written, for example, using PERL, simple queries would run several seconds more quickly.
- The statelessness of WWW Entrez requires that the Entrez database be re-initialized each time a URL is fetched. This adds approximately one second to each query. Furthermore, when several rounds of neighboring have been performed, recomputation is necessary. This computation is specific to WWW Entrez and is not performed by Entrez/Network Entrez. Again, this is due to the statelessness of the WWW browser/server interaction.
- More bandwidth is required to perform analogous operations on WWW Entrez, since Network Entrez is optimized to transfer only the unformatted data which is required (formatting is performed by the client), while WWW Entrez must transfer formatted data.
- When retrieving a large amount of text, Network Entrez only retrieves the data surrounding the records that are currently visible, while WWW Entrez must retrieve all of the formatted data.

WWW Entrez discussion

Some of the limitations of WWW and HTML include:

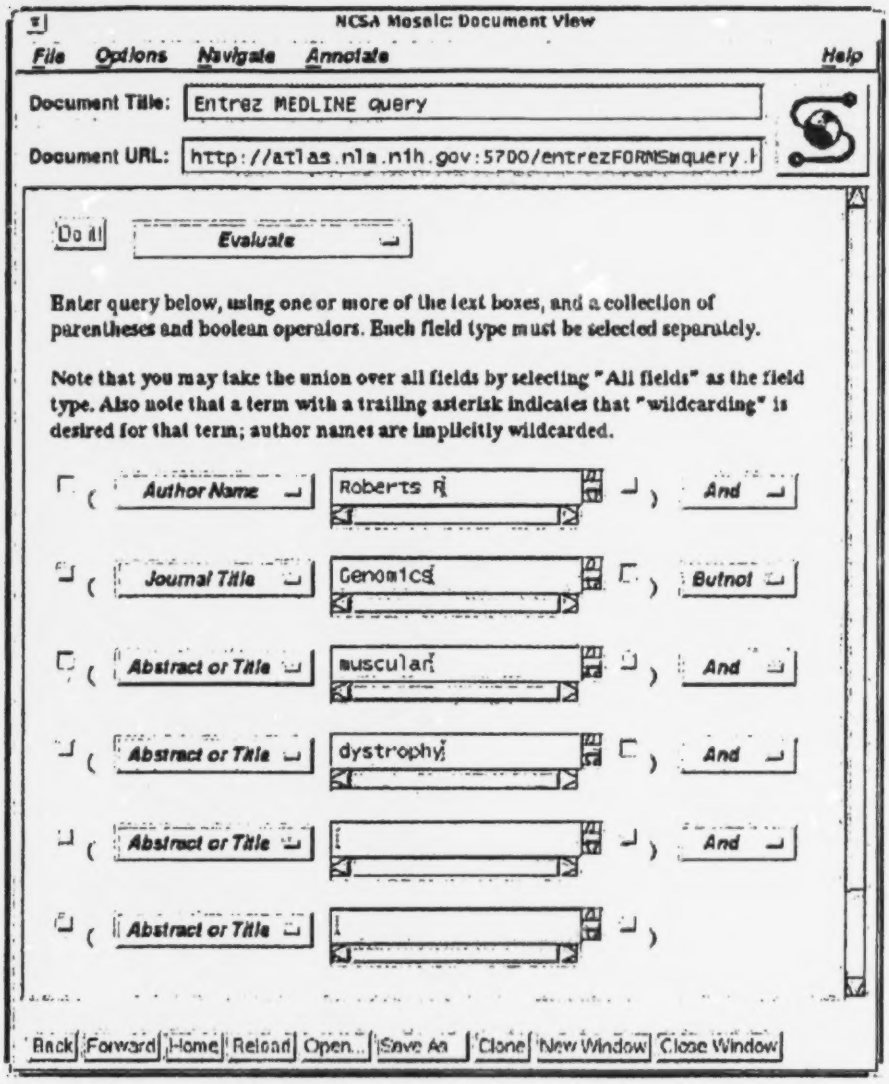
- a "click" can only result in an action when the user clicks on a piece of hypertext or presses a form submission button. Adjusting values on a FORM does not result in any action until the form submission button is pressed.
- such a click results in a new page being fetched, rather than modifications to the current display
- display of a large document in an on-demand fashion is impossible; the document must be fetched in its entirety.

In addition to these constraints, clients with FORMS capability were rare in the fall of 1993, when the WWW Entrez server was written. Even at the time of this writing (September 1994), reliable FORMS-capable WWW browsers are not ubiquitous. Therefore, it is desirable to have servers which can take advantage of FORMS, but can still be useful to users who lack FORMS-capable browsers. WWW Entrez includes both FORMS-based and non-FORMS-based interfaces.

Given the constraints of non-FORMS HTML, the simplest approach was to decompose the Entrez query interface into many screens, each corresponding to a traditional hierarchical menu system. Thus, for example, to lookup a MEDLINE article by author name, one first selects "MEDLINE" from a top-level menu, then selects "Author Name" from a menu, and finally makes a query using the browser's search window.

Given the power of FORMS, it is possible to express a complex Boolean query on a single form. The sample MEDLINE form which appears below has considerable expressive power, allowing the Boolean composition (union, intersection, and set subtraction) of several terms, where each term may use a separate indexed field.

Because of this unified approach, some users have stated that they prefer this query interface to that provided by Entrez/Network Entrez.


NCSA Mosaic: Document View
File Options Navigate Annotate Help
Document Title: Entrez MEDLINE Query
Document URL: http://atlas.nla.nih.gov:5700/entrezFORMSquery.h
Do It!
Evaluate
Enter query below, using one or more of the text boxes, and a collection of parentheses and boolean operators. Each field type must be selected separately.
Note that you may take the union over all fields by selecting "All fields" as the field type. Also note that a term with a trailing asterisk indicates that "wildcarding" is desired for that term; author names are implicitly wildcarded.
Author Name
Roberts R
And
Journal Title
Genomics
But not
Abstract or Title
muscular
And
Abstract or Title
dystrophy
And
Abstract or Title
And
Abstract or Title
Back Forward Home Reload Open... Save As Clone New Window Close Window

As an option on the single FORM, a user may choose to "browse" from 50 of the available terms rather than typing an exact term. For example, if a user doesn't remember how to spell "adenoviridae", typing "adeno" and using the "browse" option is helpful. This is also especially useful when searching by author name. Note that this is a partial emulation of Entrez's "selection mode", described earlier.

Hypertext is also a natural way to represent the hierarchical organism taxonomy which was added to the Entrez CD-ROM in May 1994, and has been found to be particularly useful to molecular phylogenists and other systematists. [4] This feature was subsequently added to WWW Entrez.

The visual clarity of hypertext permits the coherent presentation of rich text, compensating for other disadvantages of WWW and HTML. For example, the number of neighboring MEDLINE, protein, and nucleotide articles is presented to the user in WWW Entrez, smoothing his/her access to the available data. If these counts were computed and displayed in Entrez/Network Entrez, the screen would be cluttered, and, because of the extra time required to perform the computations, Entrez's otherwise quick performance would be compromised.

The power of the entrend engine allows each Entrez query to be stateless, even though a user may perform several rounds of neighboring and linking between and within the MEDLINE, protein, and nucleotide databases. This statelessness requires the recomputation of some results which could be stored, but the simplicity and power which statelessness provides is considerable. Each WWW Entrez URL is an encoding describing an originating Boolean expression or set of unique identifiers, along with the "rounds" of neighboring and linking which have been performed.

Comparison of three versions of Entrez

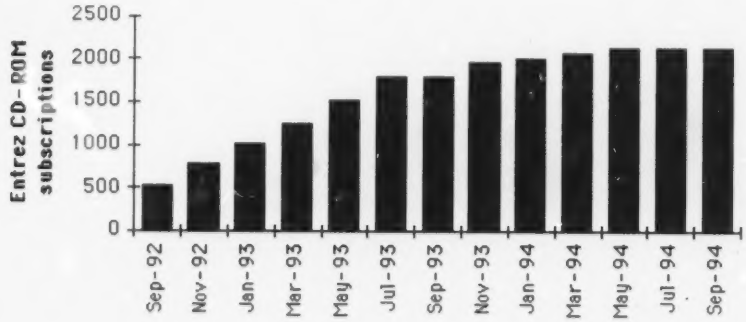
BEST COPY AVAILABLE

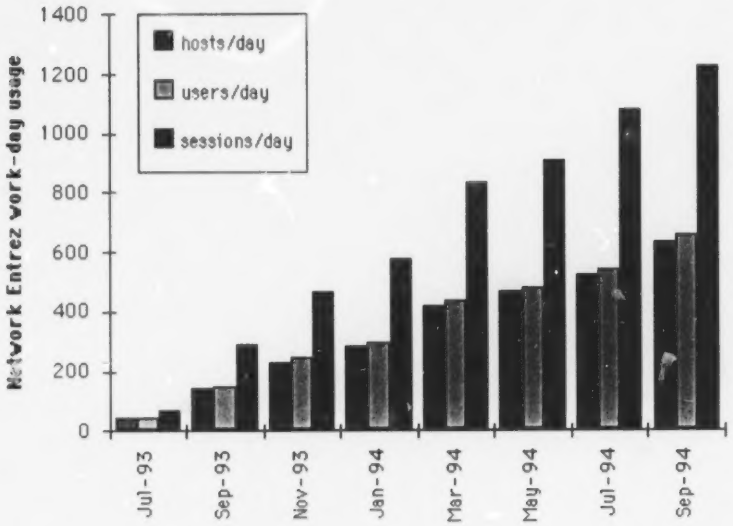
	CD-ROM	client/ server	WWW
No NCBI software required for user			√
Easily changed without changing user software			√
Usable on vt100-class terminal			√
MeSH glossary available			√
Hypertext links available to external (non-Entrez) databases			√
Weekly updates available in the future		√	√
Latest data obtainable without waiting for CD-ROM mastering and distribution		√	√
No CD-ROM drive required		√	√
Free (except for Internet operations cost)		√	√
Usable on many different hardware/software platforms	√	√	√
No registration required	√		√
"BigMed" dataset available with 6 times more MEDLINE data		√	
"on-the-fly" MEDLINE neighboring available in the future		√	
Fast for users with good network connections		√	
Structured data pre-parsed for custom-written applications	√	√	
Efficient fetching strategy when viewing a large number of abstracts	√	√	
100% supported by NCBI (non-experimental)	√	√	
"Best" ergonomic interface	√	√	
Brute-force search of data possible in custom-written applications	√		
Immune from network failures	√		

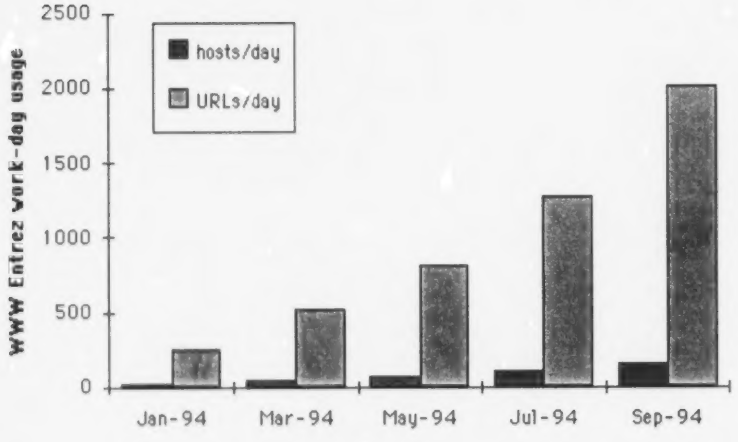
Usage

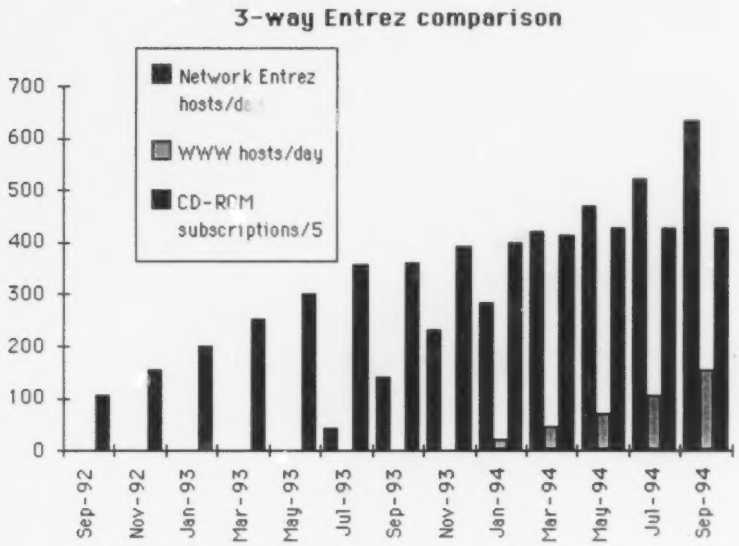
The usage of WWW Entrez and Network Entrez has grown dramatically during 1994, while the growth of the Entrez CD-ROM subscriptions has reached a plateau.

Comparing usage of the three types of Entrez is complex because there is no way to know how many users are actually using the Entrez CD-ROM and its derivative hard-disk copies. Furthermore, there is no exact analogy between Network Entrez sessions and WWW Entrez URLs. However, a more detailed study suggests that an average Network Entrez session corresponds to roughly eight WWW Entrez URLs.




Network Entrez work-day usage
1400
1200
1000
800
600
400
200
0
hosts/day
users/day
sessions/day
Jul-93
Sep-93
Nov-93
Jan-94
Mar-94
May-94
Jul-94
Sep-94


WWW Entrez work-day usage
2500
2000
1500
1000
500
0
hosts/day
URLs/day
Jan-94
Mar-94
May-94
Jul-94
Sep-94



Conclusions and future directions

The rapid growth in use of all three types of Entrez shows that it is a powerful tool which meets the needs of the molecular biologist.

In recent months, the plateau in Entrez CD-ROM subscriptions has been triggered by the wide availability of the two Internet-based versions. In some ways, this is not a bad thing, especially because the Entrez CD-ROM subscription expanded to two CDs in October 1993, and must expand to three CDs in October 1994, due to the dramatic growth in molecular sequence data.

WWW Entrez and Network Entrez complement, rather than compete with one another. WWW Entrez is useful for those who prefer a single software tool and who can accept the slower performance. Network Entrez is critical for high performance, smoother ergonomics, and custom-written applications which need pre-parsed data. Finally, for some users either Network Entrez or WWW Entrez fails to function for some technical reason. In these cases, the NCBI staff has usually been able to refer the user to the alternate Internet-based solution.

The WWW Entrez server is able to profit from the interconnection capabilities of the World Wide Web. First of all, Entrez's own data is highly interconnected. Secondly, for some protein data where external information is available on the Web, WWW Entrez points to the external Expasy, "Molecules R Us" servers. Finally, several servers point to the WWW Entrez server to obtain data, notably Expasy and the Baylor College of Medicine's sequence annotation server. The latter is especially interesting because it provides a way to annotate sequence database entries without modifying the original entry, and while pointing to the canonical Entrez sequence entry.

In the future, the Entrez WWW server will include an interface to a powerful structural molecule visualization tool (RasMol), and will contain daily-updated molecular sequence data databases.

Bibliography

1. Altschul, S. F., Gish W., Miller W., Myers E. W., & Lipman D. J. **Basic local alignment search tool.** *J Mol. Biol.* **213**, 403-10 (1990)
2. Benson D., Lipman D. J. & Ostell J. **GenBank.** *Nucleic Acids Res.* **21**, 2963-5 (1993)
3. Boguski, M. S. **Bioinformatics.** *Current Opinion in Genetics and Development* **4**, 383-388 (1994)
4. Hillis, D. M. **Phylogenetic Searching of Molecular Data Bases.** *Syst. Biol.* . **43(3)**, 461-463 (1994)
5. ISO. Information Processing - Open Systems Interconnection - *Specification of Abstract Syntax Notation One (ASN.1)* . ISO International Standard 8824, (1987)
6. Wilbur, W. J. & Coffee L. **The effectiveness of document neighboring in search enhancement.** *Inf. Process Manage.* **30**, 253-66 (1994)

Biographies

Jonathan Epstein received bachelor's and master's degrees in Computer Science from the Universities of Maryland and Texas. He spent five years developing software for real-time communications systems and network management systems, especially for satellite-based VSAT networks. For the last several years he has worked as a contractor, developing network applications software at the National Center for Biotechnology Information (NCBI).

Jonathan Kans is a Research Associate at the National Center for Biotechnology Information, a division of the National Library of Medicine at NIH. He received his A.B. in biological sciences, S.M. in immunology, and Ph.D. in genetics from the University of Chicago, and was a postdoctoral fellow at the University of California at Berkeley. His past research has centered around recombination and gene rearrangements, and, as with most of the NCBI staff, he has a strong interest in applying computers to problems of biological importance. He is the developer of the VIBRANT portable user interface, is one of the principal authors of the Entrez data retrieval application, and is the developer of the Sequin direct submission program.

Gregory Schuler received his B.S. in Biochemistry & Microbiology from the University of Maryland and his Ph.D. in Molecular Biology from Princeton University. He has performed several years of laboratory research on the role of oncogene and growth factor expression in the processes of cell differentiation and transformation. For the past several years at the NCBI, he has worked on various aspects of biological sequence analysis, such as querying and searching sequence databases and performing multiple alignments on groups of related sequences.

Contact: Jonathan Epstein *(epstein@ncbi.nlm.nih.gov)*

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: WWW Entrez: A Hypertext Retrieval Tool for Molecular Biology	
Author(s): Jonathan A. Epstein, Jonathan A. Kans, Gregory D. Schuler	
Corporate Source:	Publication Date:October 1994

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY X TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	**Level 2A**	**Level 2B**
✗	✗	✗
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature: Jonathan A. Epstein	Printed Name/Position/Title: Jonathan A. Epstein, Head, Unit on Biologic Computation	
Organization/Address: 31 Center Drive, Room 2A47, NICHD, Bethesda MD 20894	Telephone: 301-402-4563	FAX: 301-480-2336
	E-Mail Address: Jonathan_Epstein@nih.gov	Date: 19 March 2002